Motivation for Better Knowledge Distillation

Representation Gap in KD
- Challenge of the teacher-student model capacity gap.
- Existing methods are often complex and task-specific.

Noise in Distillation Features
- Student features are noisier due to the limited capacity.
- The noise leads to suboptimal distillation and performance.

Our DiffKD Approach
- A novel method using diffusion models for denoising the student features.
- Distillation on denoised student features with simple losses such as MSE.

Innovations
- Lightweight diffusion model with linear autoencoder.
- Adaptive noise matching for precise denoising.

Effectiveness
- Applicable to various feature types.
- Superior performance in multiple tasks and settings.

Method

Simultaneous Optimization with
- Task loss for training the student.
- KD loss for training the student & noise adapter.
- Diffusion loss for training the diffusion model.
- Reconstruction loss for training the linear autoencoder.

Diffusion Model
- A lightweight model with ResNet Bottleneck blocks.
- Trained with teacher features.
- Leveraged for denoising student features.

Noise Adapter
- Addresses the challenge of inexact noisy levels in student features.
- Measures the noisy level of feature.
- Complements additional Gaussian noise to feature to match the noisy level.

Experiments

ImageNet

ImageNet with Stronger Teachers

COCO

Cityscapes

Visualizations